Estimating a frequency unseen

Albers, C J; De Roos, G Th; Schaafsma, W

Published in:
Statistica Neerlandica

DOI:
10.1111/j.1467-9574.2005.00293.x

Publication date:
2005

Statistica Neerlandica (2005) Vol. 59, nr. 4, pp. 397–413

Estimating a frequency unseen: an application to ornithology

C. J. Albers*

Groningen Bioinformatics Centre, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands

G. Th. De Roos

Dorpsstraat 198, 8899 AP Vlieland, The Netherlands

W. Schaafsma

Department of Mathematics, University of Groningen

The second author is involved in a capture–mark–recapture study of some wader species. Part of his program deals with resight observations. On a particular day he visually inspects a fairly stable population to identify the ringed birds by reading their ring-number. Some ringed birds will be missed, so observations are repeated on other days. The issue of main interest is whether, after some repetitions, we can be sufficiently sure that all the ringed birds in the population have been identified or, equivalently, that the frequency of unseen birds is zero.
Most current theory is concerned with an asymptotic setting. In our 'exact' context the emphasis is on the determination of the 'probability' that the frequency of unseen birds is zero. This issue is settled by considering the more general problem of 'estimating' the frequency of the unseen birds by providing a predictive inference in the form of a probability distribution. We develop methods of inference based on the assumption of a bird-independent probability p_i of identifying a ringed bird on day i, as well as without this assumption. In Section 5 we critically examine these approaches.

Key Words: capture–mark–recapture analysis, epistemic probabilities, distributional inference.

1 Introduction

The second author is involved in a study involving the catching, measuring, ringing and colour-ringing, counting and identifying individual *Turnstones*, a wader species belonging to the *Charadriiformes* order. In this paper we develop the theory for an interesting subproblem.

*c.j.albers@open.ac.uk

Fig. 1. Ruddy Turnstones (taken from Naumann, Naturgeschichte der Vögel Mitteleuropas, Band VIII, Table 5, Gera, 1902).

De Roos has collected and is still collecting biometric, count and moult data on Ruddy Turnstones (*Arenaria interpres*, see the drawing in Figure 1). Having their breeding habitat in the high-artic tundras of Siberia and Greenland, these waders may migrate via the Frisian island Vlieland. This provides good opportunities for De Roos to study them because of the large number of breakwaters stretching into the North Sea where these birds congregate during high tides and can be caught at night if there is a dark moon, cloudy sky, and favourable tide and wind and other conditions (De Roos, 1987). The birds caught were (in the case that they hadn't aimed been ringed) ringed and colour-ringed, sexed, classified according to age, weighed, and various aspects of size and shape were measured and recorded.

During the day the birds present in the group can be counted and, at high tide, can be identified by reading ring-numbers with a telescope. Unfortunately, De Roos could not be sure of identifying, on a particular day, all the ringed birds present. However, by the procedure regularly, he can feel certain that, after a certain day (day 11 in the case of Table 1), all ringed birds have been identified. A mathematical-statistical approach will be developed to characterize the degree of (un)certainty.

Table 1. A '1' denotes that the bird indicated is spotted on the day indicated. All dates are in 1992. Note that $r_1 = 2$, $r_2 = 0$, $r_3 = 2$, $r_4 = 2$, $r_5 = 3$, $r_6 = 1$, $r_7 = 2$, $r_8 = 1$, and $r_9 = 1$ (see the text). Note also that $n = 68$ identifications have been made involving $m = 14$ birds on $k = 11$ days.

Day	Date	a	b	c	d	e	f	g	h	i	j	k	l	m	n	Total
1	20–8		1		1								1			$s_1 = 3$
2	23–8		1										1		1	$s_2 = 3$
3	26–8		1		1					1			1	1		$s_3 = 5$
4	29–8									1	1	1	1	1		$s_4 = 5$
5	2–9			1				1		1	1	1	1			$s_5 = 6$
6	5–9		1	1	1					1		1	1			$s_6 = 6$
7	8–9		1	1	1	1	1	1				1	1			$s_7 = 8$
8	11–9		1	1	1			1		1			1	1	1	$s_8 = 8$
9	14–9			1	1	1	1	1		1	1	1		1		$s_9 = 9$
10	17–9	1	1				1		1	1				1		$s_{10} = 6$
11	20–9		1	1	1	1	1	1			1		1		1	$s_{11} = 9$
Total		1	8	6	7	3	4	5	1	7	4	5	9	5	3	68

But first some notations will be defined. Using the letters a, b, . . . , n to indicate the birds identified, the data appear in the form presented in Table 1. Apart from s_i, the number of birds seen on day i ($i = 1, \ldots, k = 11$), we use the notation r_h for the number of birds identified on (exactly) h different days. Note that r_0 is the unknown value of interest: the number of ringed birds that were not seen. Finally we introduce $m = r_1 + \cdots + r_9 = 14$ as the total number of birds seen, $r = r_0 + m$ as the total number of ringed birds, and $n = \sum_{i=1}^{11} s_i = \sum_{h=1}^{11} hr_h = 68$ as the number of identifications made. Note that it takes until day 10 before each one of these 14 birds has been seen.

The mathematical-statistical problem

Given the set of data 'x' presented in Table 1, in particular the outcomes $r_1, \ldots, r_9$, we wish to construct a method of inference Q specifying a distributional inference $Q(x)$ about the number $y = r_0$ of ringed birds present but not identified by the ornithologist. In particular an assessment of the probability that $r_0 = 0$, i.e. all ringed birds have been seen, is required.

Note that the term 'distributional inference' is nothing but a new name for an old subject. It refers to the quantification of (un)certainty, belief, etc., by using probabilistic terminology. This term was introduced by KROESE *et al.* (1995), because similar terms such as Bayesian inference, fiducial inference, etc., are too closely associated to a particular methodology of generating such distributional inferences (and probability statements). With respect to the problem of interest a distributional inference $Q(x)$ about $y = r_0$ is nothing but a *concrete* probability distribution on $\{0, 1, 2, \ldots\}$ expressing an opinion about y. The name distributional inference is also used to refer to the science/technology/methodology of generating concrete distributional inferences like $Q(x)$. In our approach to distributional inference (see, e.g., KARDAUN and SCHAAFSMA, 2003) requirements of probabilistic coherency are questioned. This implies that we are somewhat critical with respect to the Bayesian approach and that we do not believe that, in the present context, the 'most reasonable' distributional inference $Q(x)$ about $y = r_0$ and the 'most reasonable' assessment $\alpha(x)$ of the probability that $y = 0$ should, necessarily, be related by $\{Q(x)\}(\{0\}) = \alpha(x)$. In this paper, however, such subtleties will be ignored.

Our problem is related to the proofreaders problem studied in the literature. POLYA (1975) considered the case of two proofreaders ('days' in our context) who read, independently of each other, the same manuscript. Let $A + C$ and $B + C$ denote the number of misprints found by reader 1 and reader 2, respectively. Here C denotes the number of commonly found misprints. If M is the (unknown) total number of misprints, then $M - A - B - C$ is the number of undiscovered misprints. Polya's estimate for this number is AB/C; the statistical uncertainties involved can be derived using the δ-method. In YANG *et al.* (1982) an 'optimal stopping rule' for rereading the manuscripts is discussed. In comparison with the proofreading problem, our problem has the advantage that probabilistic assumptions are less awkward. Another difference between our problem and that of Polya is

that the assumption of a fixed underlying population is completely natural for the proofreading problem, but not for our problem (see Section 6).

At the time we developed our theory, we did not have access to the data reported in Table 1. The examples suggested to us were such that the number k of days is so small (yet larger than 2) that it is practically impossible to falsify the hypothesis of the existence of a bird-and-day-independent 'probability' to identify ringed birds when present in the group on a given observation day. The capture-mark-recapture literature (e.g. OTIS *et al.*, 1978; WHITE *et al.*, 1982; CONN *et al.*, 2004) emphasize that 'tests for equal catchability' or 'equal identifiability' should be performed. While we worried about day-effects, we somewhat overlooked the possibility of the existence of bird-effects. Sections 2 and 3 are based on the assumption that neither bird- nor day-effects exist. In Section 4 day-effects are allowed. After a discussion in Section 5, the existence of bird-effects or, more precisely, nonconstancy of the population, (obvious from Table 1) will be dealt with in Section 6.

2 A simple model

We have the outcomes $(r_1, \ldots, r_{11}) = (2, 0, 2, 2, 3, 1, 2, 1, 1, 0, 0)$ and need the latent outcome r_0 or, more precisely, the 'probabilities' that $r_0 = 0, 1, 2, \ldots$, respectively. There are, of course, many approaches to performing such extrapolation. One such approach is to assume that $r_0, r_1, \ldots, r_k$ are outcomes of independent Poisson variables with parameters λ_h satisfying some model $\lambda_h = \lambda_\theta(h)$, e.g. with $\lambda_\theta(h) = \exp(\theta_1 + \theta_2 h)$ $(h = 0, \ldots, k)$ (see, also, STAM, 1987, for an alternative approach based on negative binomial distributions). The models we shall use have a more realistic appearance. In this and in the next section we assume (1) that the population is constant (actually, the total number of birds, ringed and unringed, varies between 60 (on day 1) and 72 (on day 10)) and (2) that there is a fixed (unknown) probability p that the j-th ringed bird is seen on day i. Here $j = 1, \ldots, r = \sum_{h=0}^{k} r_h$ and $i = 1, \ldots, k$ $(=11)$. Note that $r = r_0 + m = r_0 + 14$. Making some independence assumptions in addition, the essence of Table 1 is captured in the Kolmogorovian setting $(\Omega, \mathcal{F}, \mathrm{P})$ where Ω is the space of all $k \times r$ matrices ω with

$$\omega_{i,j} = \begin{cases} 1 & \text{if bird } j \text{ is seen on day } i \\ 0 & \text{otherwise} \end{cases}$$

and the probability of each matrix ω is

$$\mathrm{P}(\{\omega\}) = p^n (1-p)^{kr-n}$$

where, as indicated before, $n = \sum_{h=1}^{k} h r_h = \sum_{i=1}^{k} s_i = 68$ is the total number of identifications made. Note that the theoretical maximum kr for n appears if all r $(r \geq 14)$ birds are seen on all $k = 11$ days. The random variables (some of these are 'statistics', in the sense that their outcome is available) we are interested in are defined by

1.	$S_i(\omega) = \sum_{j=1}^{r} \omega_{i,j}$	the number of birds identified on day i
2.	$T_j(\omega) = \sum_{i=1}^{k} \omega_{i,j}$	the number of times bird j is seen
3.	$R_h = \#\{j \mid T_j = h\}$	the number of birds seen on h out of k days
4.	$M = \sum_{h=1}^{k} R_h$	the number of birds seen at least once
5.	$N = \sum_{h=1}^{k} hR_h = \sum_{i=1}^{k} S_i$	the total number of identifications made.

The following statements are trivial.

1. $S_1, \ldots, S_k$ are i.i.d., $S_i \sim B(r, p)$
2. $T_1, \ldots, T_r$ are i.i.d., $T_j \sim B(k, p)$
3. $(R_0, \ldots, R_k) \sim \text{Multinomial}(r; (1-p)^k, \binom{k}{1} p(1-p)^{k-1}, \ldots, p^k)$
4. $M \sim B(r, 1 - (1-p)^k)$
5. $N \sim B(rk, p)$
6. Given $\{T_j \geq 1\}$, the conditional distribution of T_j is defined by the probabilities

$$\mathrm{P}(T_j = c \mid T_j \geq 1) = \frac{\binom{k}{c} p^c (1-p)^{k-c}}{1 - (1-p)^k} \quad (c = 1, \ldots, k)$$

7. The conditional distribution of N, given $\{M = m\}$ corresponds to that of the sum $\sum_{g=1}^{m} N_g$ of independent random variables $N_1, \ldots, N_m$, all having the distribution specified under 6.

3 Settling the issue under the assumptions of Section 2

There are, of course, various reasons to question the assumption that a fixed probability p exists, independently of i and j, that the j-th ringed bird is seen on day i. We worried, for example, about the possibility of a day-effect because weather conditions are likely to play a role: on one day birds might be less likely to be identified than on another. This suggests that the assumption of a fixed p is unrealistic. In this section, however, we will use the assumption of a constant probability.

It seems reasonable to concentrate our attention on the outcome $x = (m, n)$ (= (14, 68)) of (M, N) and on the number k (=11) of days (in Table 1; later we shall consider alternative situations where a distributional inference is made on the basis of, e.g., the first five observation days, see Figure 2, or the last one or three, see Figure 3). To start with, we estimate p by equating

$$\mathbf{E}(N \mid M = m) = m\mathbf{E}(T_j \mid T_j \geq 1) = mkp/(1 - (1-p)^k)$$

to the outcome n or, equivalently, by computing the estimate $\hat{p}$ as the solution of

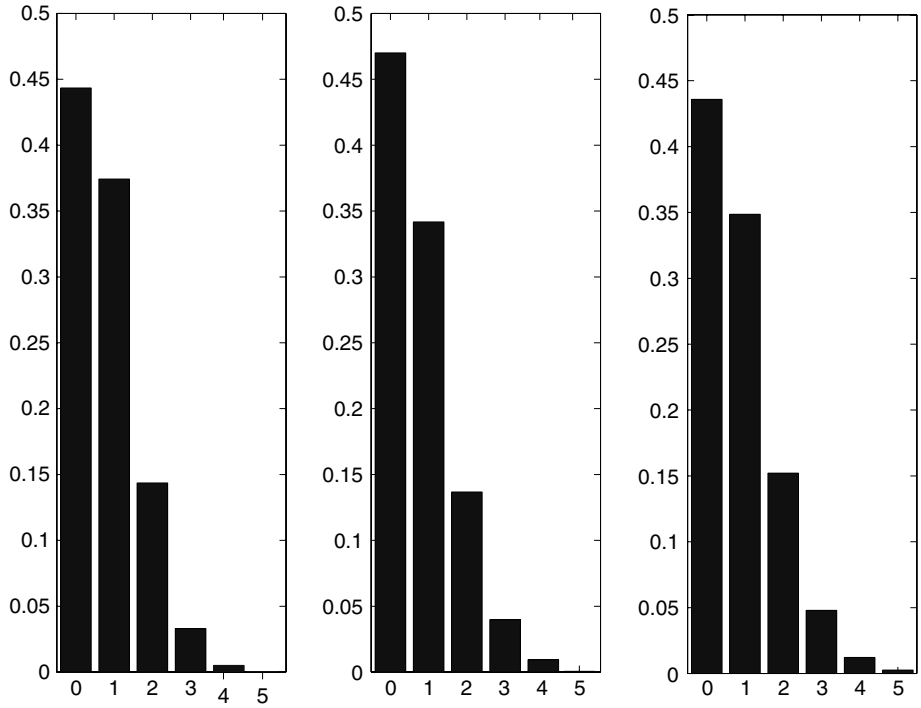

Fig. 2. Visualization for the case $k = 5$. From left to right, the inferences 'Bin$(\hat{r}, 1-q)$', $Q(m) \approx \tilde{Q}(m)$ and $Q_{\mathrm{LIK}}(m)$ about r_0 are displayed. We recommend $Q(m) \approx \tilde{Q}(m)$.

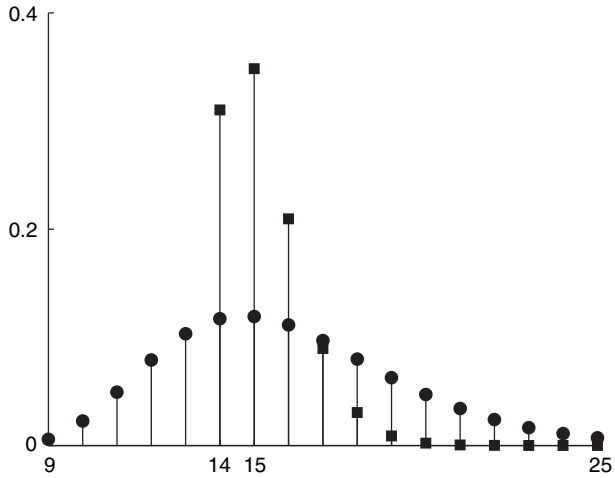

Fig. 3. Opinions about r, expressed as distributional inferences, based on the last row (circles) and on the last three rows (squares) of Table 1.

$$p = \frac{n}{mk}(1 - (1 - p)^k).$$

Next, ignoring the uncertainties involved in the estimation of p and concentrating our attention on the outcome m of

$$M \sim B(r, 1 - (1 - \hat{p})^k),$$

an assessment has to be made of the 'probability' that $r_0 = r - m$ is equal to 0 and, more generally, a distributional inference $Q = Q(m)$ has to be made specifying such probabilities for all possible values 0, 1, 2, ... of the frequency r_0 of birds that were not seen.

Classical statisticians will be attracted to the idea that such 'predictive distribution' Q can be obtained from the 'factual' result, implied by Section 2, that

$$R_0 \sim B\Big(r, (1 - p)^k\Big).$$

They will replace the unknown values r and p in this equation by certain estimates. A referee suggested using the moment estimates $\hat{r}$ and $\hat{p}$, which one can obtain by equating the outcome m of M with $\mathbf{E}M = r(1 - (1 - p)^k)$ and the outcome n of N with $\mathbf{E}\, N = rkp$. Such a plug-in approach is rather natural and a more sophisticated approach should be in line with it. Some refinement, however, cannot be dispensed with because (1) statistical uncertainties in the estimates $\hat{r}$ and $\hat{p}$ should be incorporated, (2) the estimate $\hat{r}$ is not an integer. As indicated before (see Section 1, the mathematical-statistical problem), theories of Bayesian inference, fiducial inference, etc. have been developed to cope with such refinements. Ignoring the statistical uncertainty in estimating p (i.e. by equating p and $\hat{p}$), we have to settle the following issue.

Problem. Given is the outcome m of $M \sim B(r, q)$ where

$$q = 1 - (1 - \hat{p})^k$$

and required is a distributional inference $Q(m)$ about the frequency $r_0 = (r - m)$ of unseen birds.

We shall subsequently consider a 'formal-Bayes' approach and a 'fiducial' approach. (We shall not use the 'personalist-Bayes' approach based on a 'proper prior distribution' for the number r_0 of birds unseen because we consider it unreasonable to ask De Roos to specify such a prior. The formal-Bayes approach we shall follow is not necessarily the most appropriate. In Section 4 we shall play with the idea of assigning probabilities 1/2 to the possibilities $r_0 = 0$ and $r_0 = 1$ or, in the context used there, to the possibilities $r = m$ and $r = m + 1$.)

The 'formal Bayes' approach we would like to consider here is that the distributional inference $Q = Q_{\mathrm{LIK}}(m)$ about r_0 is simply obtained by standardizing the likelihood function. Using M_θ as a notation for a (fictitious) random variable with the same distribution as M if r_0 happened to be equal to $\theta \in P = \{0, 1, \ldots\}$, we obtain from

$$\mathrm{P}(M_\theta = m) = \binom{r(\theta)}{m} q^m (1-q)^{r(\theta)-m}$$

where $r(\theta) = m + \theta$, that the likelihood function $l_m(\theta) = \mathrm{P}(M_\theta = m)$ provides the probabilities

$$l_m(\theta) / \sum_{\theta=0}^{\infty} l_m(\theta) = \binom{m+\theta}{m} q^{m+1} (1-q)^\theta$$

of $\theta = 0, 1, \ldots,$ after standardization. These are the probabilities of the distributional inference

$$Q_{\mathrm{LIK}}(m) = \mathrm{NegBin}(m+1, q).$$

Assigning prior mass 1 to the possibilities $\theta = 0, 1, \ldots$ for r_0 is very questionable, especially if the main interest is in the extreme case $\theta = 0$: that De Roos has seen all ringed birds present. (Assigning prior probabilities 1/2 to $\theta = 0$ and $\theta = 1$ is a possibility but, in our opinion, not a very attractive one either.)

The 'fiducial approach' (see KROESE *et al.*, 1995 and SALOMÉ, 1998) is based on the simple 'classical-statistical' idea that the distribution function G_m of the distributional inference $Q(m)$ that we try to construct should be determined by identifying $G_m(\theta)$ with the 'most reasonable' degree of belief $\alpha_\theta(m)$ in the truth of the hypothesis H_θ: $r_0 \leq \theta$. The next question is, of course, how to specify such a most reasonable degree of belief. In the fiducial approach it is recommended that one uses 'some' p-value. The construction of distributional inferences for real-valued unknown parameters is then reduced to the construction of p-values for a family of testing problems.
If we use the one-sided p-value

$$\alpha_\theta(m) = \mathrm{P}(M_\theta \geq m) = \sum_{s=m}^{r(\theta)} \binom{r(\theta)}{s} q^s (1-q)^{r(\theta)-s}$$

as degree of belief in H_θ: $r_0 \leq \theta$ and define $G_m(\theta) = \alpha_\theta(m)$ then, using

$$\mathrm{P}(\mathrm{Bin}(r, q) \geq m) = \mathrm{P}(\mathrm{NegBin}(m, q) \leq r - m),$$

we infer that

$$G_m(\theta) = \mathrm{P}(\mathrm{NegBin}(m, q) \leq r(\theta) - m = \theta)$$

and, hence, that the distributional inference

$$Q(m) = \mathrm{NegBin}(m, q)$$

is obtained.

In the theory of distributional inference, alternative solutions are discussed because it is not completely reasonable to identify the one-sided p-value with the degree of belief in H_θ. In some situations assigning the 'mid-p-value'

$$\alpha_\theta(m) = \frac{1}{2}\mathrm{P}(M_\theta \geq m) + \frac{1}{2}\mathrm{P}(M_\theta \geq m+1)$$

is recommendable. For the present problem, KARDAUN and SCHAAFSMA (2003) show that the distributional inference

$$\tilde{Q}(m) = \left(\frac{1}{2} + \frac{1}{2}q\right)\mathrm{NegBin}(m, q) + \left(\frac{1}{2} - \frac{1}{2}q\right)\ \mathrm{NegBin}(m+1, q)$$

provides a slight improvement (in the sense that $\tilde{Q}$ is approximately 'weakly similar', whereas Q is not).

Numerical evaluation of the data in Table 1

The above methods provide $\hat{p} = 0.441$ as the solution of $14{\cdot}11p = 68(1 - (1 - p)^{11})$ and that $Q_{\mathrm{LIK}}(m) = \mathrm{NegBin}(15, 0.998)$ assigns the probability $0.998^{15} = 0.975$ to H_0: $r_0 = 0$ while $Q(m) = \mathrm{NegBin}(14, 0.998)$ assigns the probability $\{Q(m)\}(0) = 0.998^{14} = 0.977$. Note that $Q(m) \approx \tilde{Q}(m)$ because $q \approx 1$.

Of course, also after the first, second, ..., etc., day, De Roos could have tried to make a distributional inference about the number of unseen birds. Figure 2 displays, for $k = 5$, the distributional inferences $\mathrm{Bin}(\hat{r}, 1-q)$, $Q(m) \approx \tilde{Q}(m)$ and $Q_{\mathrm{LIK}}(m)$ about r_0. As stated earlier, $\hat{r}$ (=10.78) is not an integer; the displayed inference is based on Bin(10.78, q): = 0.22Bin(10, q) + 0.78Bin(11, q). Although, as suggested, the right-tails of $\tilde{Q}$ and Q_{LIK} are heavier than that of $\mathrm{Bin}(\hat{r}, 1-q)$, the differences between the three inferences are practically ignorable in this case $k = 5$. This, however, will not be true in general. We consider the distributional inference $\tilde{Q}(m)$ as most reasonable because of the underlying theory but we will not object if somebody else proposes $Q(m)$. With respect to Q_{LIK} we are less positive because, if one is interested in the truth or falsity of H_0: $r_0 \leq \theta$, then it does not seem reasonable to assign prior measure $\theta + 1$ to H_0 and prior measure ∞ to A_θ: $r_0 \geq \theta + 1$.

The outcomes (k, m_k, n_k), the estimates $\hat{p} = \hat{p}_k$, as well as the assessments $\alpha_0(m) = \{Q(m)\}(\{0\})$ and $\tilde{\alpha}_0(m) = \{\tilde{Q}(m)\}(\{0\})$ are displayed in Table 2. It is not until after the ninth day that we are sufficiently certain (at $\alpha = 5\%$) that De Roos has identified all the ringed birds present. Nevertheless the next day (day 10), two new birds appeared (birds k and h).

Initially, we (Albers antd Schaafsma) were completely satisfied by this approach, the 'fiducial' inferences $Q(m)$ and $\tilde{Q}(m)$ in particular. However, after discussion with

Table 2. First three rows: number, m_k, of birds spotted at least once and total number, n_k, of observations, both after k days. Last three rows: the estimate $\hat{p}$ of p and the probabilities $\{Q(x)\}(0)$ and $\{\tilde{Q}(x)\}(0)$ assigned to the event that all ringed birds present in the group were identified after day k.

k	2	3	4	5	6	7	8	9	10	11
m_k	4	6	8	10	10	12	12	12	14	14
n_k	6	11	16	22	28	36	44	53	59	68
$\hat{p}$	0.667	0.559	0.456	0.408	0.454	0.419	0.455	0.490	0.420	0.441
$\{Q(x)\}(0)$	0.624	0.583	0.481	0.470	0.765	0.762	0.910	0.972	0.941	0.977
$\{\tilde{Q}(x)\}(0)$	0.620	0.581	0.479	0.469	0.765	0.762	0.910	0.972	0.941	0.977

De Roos and other ornithologists and, especially, after seeing Table 1, we lost the conviction that the theories of Sections 2 and 3 are satisfactory. The existence of a fixed probability p, independent of day and bird, is obviously questionable. Another (minor) drawback is that the statistical uncertainties involved in $\hat{p}$ have been ignored. In the next section, a theory will be presented to deal with the case that day-effects are allowed.

4 Taking day effects into account

To adapt the theory of the previous sections to the situation where on day i a probability p_i is involved, we shall condition on $\{S_1 = s_1, \ldots, S_k = s_k\}$ to get rid of $p_1, \ldots, p_k$. This is convenient and natural, but not compelling because $s_1, \ldots, s_k$ contains some information about r, e.g. that $r \geq s = \max(s_1, \ldots, s_k)$. Assuming that bird-effects are absent, we have that, given the number s_i of ringed birds on day i, all $\binom{r}{s_i}$ combinations of s_i birds from the r ringed birds available have the same probability $1/\binom{r}{s_i}$ of being the s_i ones seen. Following a recommendation by A.J. Stam (personal communication), Ω consists of points ω which are composed of the sets

$$\{v_{i,1}, \ldots, v_{i,s_i}\} \subset \{1, \ldots, r\}$$

of the identification or, rather, index numbers of the s_i birds seen on day i ($i = 1, \ldots, k$). Assuming independence, the conditional probability distribution P on $(\Omega, \mathcal{F})$ is determined by

$$\mathrm{P}(\{\omega\}) = \begin{cases} 1/\left(\binom{r}{s_1}\binom{r}{s_2}\cdots\binom{r}{s_k}\right) & \text{if } S_i(\omega) = s_i (i = 1, \ldots, k) \\ 0 & \text{otherwise.} \end{cases}$$

Note that here, and elsewhere, we use notations for random variables, conceptually defined in Section 2, but now defined on a different Ω-space such that, e.g., S_i is not 'random' at all. As defined earlier, T_j denotes the number of times bird j has been observed. The (conditional) distribution of the relevant observable $M = \sum_{j=1}^{r} \mathbf{1}_{\{T_j \geq 1\}} = r - R_0$ can now be studied for any a priori possible value $\theta \in P = \{s, s+1, \ldots\}$ of r, where, as indicated before, $s = \max(s_1, s_2, \ldots)$. Standard theory (see, e.g., PARZEN, 1960, Section 2.6), provides that

$$\mathrm{P}(M = m) = \sum_{j=m}^{r} (-1)^{j-m} \binom{j}{m} Q_j$$

and that

$$\mathrm{P}(M \geq m) = \sum_{j=m}^{r} (-1)^{j-m} \binom{j-1}{m-1} Q_j$$

where

$$Q_0 = 1$$

$$Q_1 = \sum_{j=1}^{r} \mathrm{P}(A_j)$$

$$Q_2 = \sum_{j_1=1}^{r} \sum_{j_2=j_1+1}^{r} \mathrm{P}(A_{j_1} A_{j_2})$$

$$\vdots$$

$$Q_r = \mathrm{P}(A_1 A_2 \ldots A_r)$$

and $A_j = \{T_j \geq 1\}$ $(j = 1, \ldots, r)$. The (conditional) probability that bird j is not seen on day i is equal to $1 - (s_i/r)$. Hence, the probability that bird j is never seen is equal to

$$\mathrm{P}(T_j = 0) = \prod_{i=1}^{k} \Big(1 - \frac{s_i}{r}\Big).$$

As r is unknown, we introduce the auxiliary random variable M_θ such that M_θ has the distribution which M would have had if $r = \theta$. The result just mentioned provides the relevant 'physical' probabilities $p_\theta(\mu) = \mathrm{P}(M_\theta = \mu)$ $(\mu = s, s + 1, \ldots, \theta)$ whether or not a priori probabilities are specified. We are interested in the construction of (epistemic) posterior probabilities

$$q_m(\theta) \quad (\theta = m, m + 1, \ldots)$$

specifying the opinion we should have about r after observing the outcome m of M. The posterior probability $q_m(m)$ is of particular interest because it refers to the (epistemic) probability that De Roos has identified all ringed birds in the population. Two issues are involved:

1. The determination of $p_\theta(\mu)$
2. How to convert the $p_\theta(\mu)$, given the outcome m of M, into the $q_m(\theta)$.

The distribution of M has been derived in the above, at least in principle. Its expectation is given by

$$\mathbf{E}(M) = \sum_{j=1}^{r} \mathrm{P}(T_j \geq 1) = \sum_{j=1}^{r} \big(1 - \mathrm{P}(T_j = 0)\big) = r - r \prod_{i=1}^{k} \Big(1 - \frac{s_i}{r}\Big).$$

Thus having obtained

$$\mathbf{E}(M_\theta) = \theta \left(1 - \prod_{i=1}^{k} \Big(1 - \frac{s_i}{\theta}\Big)\right)$$

we shall content ourselves by providing approximate $p_\theta(\mu)$'s by equating $\mathcal{L}(M_\theta)$ to the distribution on $\{s, s + 1, \ldots, \theta\}$ which maximizes the entropy

$$-\sum_{\mu=s}^{\theta} p(\mu)\log p(\mu)$$

under the restrictions

$$p(\mu)\geq 0,\quad \sum_{\mu=s}^{\theta} p(\mu)=1,\quad \sum_{\mu=s}^{\theta}\mu p(\mu)=\mathbf{E}(M_\theta).$$

This maximum entropy approach (cf. JAYNES, 2003) provides

$$\tilde{p_\theta}(\mu)=\exp(c_\theta\mu-\psi(\theta))\ (\mu=s,\ldots,\theta)$$

with $\psi(\theta)=\log\sum_{\mu=s}^{\theta}\exp(c_\theta\mu)$ and with c_θ such that $\sum\mu\tilde{p_\theta}(\mu)=\mathbf{E}\,M_\theta$. We believe that it is not very reasonable, in the context of Table 1 with $k=11, s=9, m=14, n=68$, to convert the $\tilde{p_\theta}(\mu)$ into posterior probabilities $q_m(\theta)$ by normalizing the likelihood function $l_m(\theta)=\tilde{p_\theta}(m)$ or, equivalently, by using the formal Bayes approach with improper prior $w(\theta)=1(\theta=s,\ s+1,\ \ldots)$. In our opinion, it is more reasonable to use some form of Fisher's fiducial argument, e.g. that where the distribution function

$$G_m(z)=\sum_{\theta=m}^{z} q_m(\theta)$$

of the distributional inference about r is equated to the p-value $\alpha_z(m)=\mathrm{P}(M_z\geq m)$ or to the symmetrized p-value $\tilde{\alpha_z}(m)=1/2\mathrm{P}(M_z\geq m)+1/2\mathrm{P}(M_z\geq m+1)$ where, of course, the approximate values

$$\mathrm{P}(M_z\geq m)\approx\sum_{\mu=m}^{z}\tilde{p_z}(m)=\mathrm{e}^{-\psi(z)}\sum_{\mu=m}^{z}\mathrm{e}^{c_z\mu}=\mathrm{e}^{-\psi(z)}(\mathrm{e}^{c_z(z+1)}-\mathrm{e}^{c_z(m)})/(1-\mathrm{e}^{c_z})$$

are used. To avoid extensive computations and, also, for some other reasons (see Section 5), we shall proceed somewhat differently. With respect to the determination of $q_m(m)$, we consider it appropriate, in the present context, to use a Bayesian approach where the data-dependent prior $w(\theta)=1/2\quad(\theta=m,\ m+1)$ is used. It provides us with

$$q_m(m)=\frac{p_m(m)}{p_m(m)+p_{m+1}(m)}\approx\ (1+\mathrm{e}^{(c_{m+1}-c_m)m-\psi(m+1)+\psi(m)})^{-1}$$

If, in the situation of Table 1, the question is considered whether all ringed birds were identified in the first 11 days, then $s=\max(s_i)=9, k=11$, and $m=14$. Taking $\theta=14$ provides $\mathbf{E}\ M_{14}=14(1-\prod(1-s_i/14))=13.984$ such that the maximum entropy solution satisfies $\tilde{p}_{14}(\mu)\propto\exp(c_{14}\mu)$ with $c_{14}=4.175$. Hence

$$(\tilde{p}_{14}(9),\ldots,\tilde{p}_{14}(14))=(0,0,0,0,0.015,0.985).$$

For the denominator of $q_m(m)$, given above, it is also necessary to look at the situation $\theta=m+1$. This provides $\mathbf{E}\ M_{15}=14.968$, and $\tilde{p}_{15}(\mu)\propto\exp(c_{15}\mu)$ with $c_{15}=3.470$. Hence

Table 3. First three rows: number of birds spotted at least once (m) at day k and maximum number $s = \max(s_1, \ldots, s_k)$ of observations per day until day k. Last row: probability that all ringed birds have been seen after day k.

k	2	3	4	5	6	7	8	9	10	11
m	4	6	8	10	10	12	12	12	14	14
s	3	5	5	6	6	8	8	9	9	9
$\tilde{p}_m(m)$	0.750	0.750	0.681	0.664	0.836	0.839	0.940	0.984	0.958	0.985
$\tilde{p}_{m+1}(m)$	0.323	0.302	0.256	0.250	0.195	0.186	0.099	0.036	0.069	0.030
$q_m(m)$	0.700	0.713	0.727	0.726	0.811	0.819	0.905	0.965	0.933	0.970

$$(\tilde{p}_{15}(9), \ldots, \tilde{p}_{15}(15)) = (0, 0, 0, 0, 0.001, 0.030, 0.969).$$

As a consequence we obtain the (approximate) probability

$$q_{14}(14) = \frac{\tilde{p}_{14}(14)}{\tilde{p}_{14}(14) + \tilde{p}_{15}(14)} = \frac{0.985}{0.985 + 0.030} = 0.970,$$

that $r = 14$ and, hence, $r_0 = 0$. According to this model, the hypothesis that at least one ringed bird has not been seen after day 11 is rejected at $\alpha = 5\%$ (or, almost equivalently, the hypothesis that all birds have been seen is accepted). Table 3 provides such 'probabilities' $q_m(m)$ in favor of H_0: $r_0 = 0$, after inspection day k. The agreement with the results in Table 2 is rather satisfying.

The theory in this section is not completely compelling, since much more information is needed about the accuracy of the maximum-entropy approximation to the true distribution $\mathcal{L}(M_\theta)$ and, also, since the data-dependent prior $w(m) = w(m+1) = 1/2$ is very questionable if k is smaller than, say, 9. (Elaboration along the lines of the fiducial argument has not been performed because of the awkward conclusion of the next section.)

5 Discussion

As indicated at the end of Section 1, the theory was developed without having access to the real data. The attention was concentrated on a small number of consecutive days (say $k = 6$) such that falsification of the hypothesis of bird-independent experimental probabilities would not be feasible. In such situations the research worker may decide to make the assumption of 'no bird-effects'. If the hypothesis $p_1 = \cdots = p_k$ of 'no day-effects' is acceptable (as in Table 1), then the theory of Section 3 is applicable. If this hypothesis is not reasonable, then one might use the theory of Section 4. That the results reported in Tables 2 and 3 are not much different could have been expected on the basis of the acceptability of the hypothesis $p_1 = \cdots = p_k$. In practice, it may very well happen that day-effects are present. Ornithologist J.B. Hulscher was dealing with counting all ringed Oystercatchers (*Haematopus ostralegus*) on Schiermonnikoog (another Frisian island). In his

experience (personal communication) the frequencies $s_1, \ldots, s_k$ of birds counted on k consecutive days were too different to assume a common p. This implies that the theory in Section 4 may be of practical interest as well.
However, if we study the frequencies $(r_1, \ldots, r_{11}) = (2, 0, 2, \ldots, 0)$ of Turnstones with $1, 2, \ldots, 11$ identifications, then these frequencies are in obvious conflict with the probabilistic assumption of 'no bird-effects'. This leaves us with an awkward issue. We have applied the theory developed in Sections 2, 3 and 4 to a table which is in conflict with the assumption of 'equal watchability'. In less extensive applications this assumption may be acceptable, but for Table 1 our theory is 'dead'.

6 Life after death?

In the previous sections we ignored the fact that our observations were not from a fixed population, but that the population was slowly expanding in time (especially around the sixth day), see Table 4. If we look at Table 1, we get the impression that, e.g., birds a and h were not present in the population in the beginning. This explains, at least partly, that the frequencies $(r_1, \ldots, r_{11}) = (2, 0, 2, 2, 3, 1, 2, 1, 1, 0, 0)$ are in contrast with the hypothesis of a fixed probability. It also explains why the estimates $\hat{p}_k$ in Table 2 display a decreasing trend. Fortunately, we can exploit the following information, obtained from Table 1. If a bird is seen at least twice, then there is a first and a last day, and a number of days in between. Restricting the attention to the birds with, at least, one intermediate observation (birds b, c, etc. indicated in Table 5) we can count the number u_i of such intermediate observation days ($u_i = 9$ for bird b, see Table 1) and the number v_i of these u_i days ($v_i = 6$ for bird b, see Table 1) on which the bird was identified. Table 5 provides the basis for the following argument. If we assume that rejection of the hypothesis of a fixed probability is entirely due to the (obvious) fact that some birds are not always present, then we can concentrate the attention on the probability of seeing a ringed bird, *when present* in the population on any day. Ignoring day-effects and assuming

Table 4. Total number of birds on the breakwaters each day.

day	1	2	3	4	5	6	7	8	9	10	11
population	60	63	61	62	60	65	71	70	71	72	70

Table 5. Days between first and last sighting, and number of in-between sightings, given for birds that are identified at least twice.

bird i	b	c	d	e	f	g	i	j	k	l	m	n	total
u_i	9	5	9	3	3	5	6	6	4	9	6	8	73
v_i	6	4	5	1	2	3	5	2	3	7	3	1	42

that birds seen at least twice were present on *all* intermediate days, we can use Table 5 to compute the relative frequency $42/73 = 57.5\%$ as an estimate of the probability of reading the ring-number of a bird *if it is present* on a specific day. Table 1 suggests (see birds j, m and, especially, n) that birds are not necessarily present on *all* intermediate days. That is why the denominator in 42/73 is too large and the statement should be that if, on some day, a ringed bird is present in the population, then De Roos will identify its ring number with probability *at least* 57.5%. We were, obviously, not sufficiently precise with respect to the population concept. Moreover, we have ignored some, possibly relevant, information about colour-rings.

With respect to the *population-concept*, a general methodological perspective is based on the idea that there is a gradual increase of concreteness – and decrease of abstractness – if one considers the sequence epistemology/mathematical statistics/applied statistics, /actual scientific research. An example about birds, from epistemology, is HEMPEL's ravens paradox (1965). The essence is as follows. Sitting by his desk is a rational man, looking around his study and seeing books, pictures, the cat, etc., and the curtains closed in from of the windows, he notices that every single non-black item he sees is not a raven. But if 'non-black' implies 'non-raven' then, by inversion, all ravens must be black. The flaw in this argument is, of course, that the 'universe of discourse' is tacitly extended from the factual population of things in the rational man's study to the entire world. Hempel concludes that *armchair ornithology should not give us beliefs about real birds in the wild*. This is in line with the conclusion in the above that we should have been more precise about the 'universe of discourse'. With respect to his population, De Roos was very clear: it consists of all birds present on the high-tide roosting site where they (tend to) congregate during high-tides during the day (some birds, but not too many, may occasionally be elsewhere on Vlieland). This population of birds was fairly, but not exactly, constant during the inspection period considered: it is expanding somewhat (see Table 4) because of delayed arrivals.

With respect to the existence of *additional information*, De Roos did not only denote the ring-number, he also read the colour-rings whenever possible. For some birds the colour-ring was read but the identification number was not, because the bird flew away before identification was successful. Data presented in Table 6

Table 6. Extension of Table 4. On day i, De Roos read the colour code of the rings of c_i birds, and for s_i birds, the ring-number was also noted. Note that assuming the presence of 14 (or 15) birds during days 7 to 10, the probability of reading the identification number is (about) 57% (or 53%).

day	i	1	2	3	4	5	6	7	8	9	10	11
population	n_i	60	63	61	62	60	65	71	70	71	72	70
colour-rings	c_i	6	6	8	8	10	9	10	14	13	14	14
identifications	s_i	3	3	5	5	6	6	8	8	9	6	9

indicate that the c_i are increasing more rapidly than the s_i. The explanation is that during the earlier days of the inspection period it is more difficult to approach the population such that colour-rings and, especially, ring-numbers can be identified: the birds had to become accustomed to their new environment and to the ornithologist's behaviour. The conclusion of the existence of a fixed probability of at least 57.5% to identify a ringed bird (if it is present) was made too hastily. Tables 1 and 6 are in line with the statements that (1) the probability of reading the colour-ring gradually increases from about 0.50 on day 1 to about 0.95 during the last 5 days; (2) the probability of reading the identification number, given the reading of the colour-ring, is fairly constant; it is about $\sum s_i / \sum c_i = 0.6$; (3) the probability of identifying the bird by reading its ring-number is about 0.57 during the last five inspection days in Table 1; (4) in principle, birds stay in the population from the day of their arrival until the day of their (common) departure which is beyond the inspection period; (5) if this is used as an assumption then Table 1 can be used to provide the estimate $54/91 = 0.59$ for the probability that a ringed bird, when present, is identified.

Concluding remarks. De Roos collected the data presented in Table 1 because he had to inspect his population. He does not want to miss any ringed bird because he is interested in making a survival analysis of Turnstones. It is fair to conclude that he should inspect his population more frequently in the second half of the period that it is present on Vlieland: first because identification is somewhat easier and second because of the late arrivals. If we restrict the attention to these later inspections, say the last k days, then we can conclude that (1) if De Roos wants to have a probability of at least 95% of identifying all ringed birds available and is expecting 15 to 20 ringed birds, then he should choose k such that

$$\left(1 - (0.43)^k\right)^{20} \geq 0.95,$$

this provides $k \geq 8$.

(2) In practice, De Roos is not inspecting his population that often. If he inspects his population in some year k times and succeeds in identifying m birds by reading their ring-number, then he will worry about the frequency $r_0 = r - m$ of unseen birds. If inspections took place in the second half of the period that these Turnstones are present, then the situation of Section 3 becomes of interest, with $\hat{p}$ replaced by 0.57.

Example. Suppose De Roos inspects his population on one day only providing $m = 9$ identifications, like on the 11th day of Table 1. Using $k = 1$, $q = 0.57$ we recommend the distributional inference

$$\tilde{Q}(m) = 0.79\text{NegBin}(9, 0.57) + 0.21\text{NegBin}(10, 0.57)$$

for the frequency of ringed birds not identified (see Section 3). By shifting the distribution $m = 9$ units to the right, we obtain a distributional inference about the total frequency r of ringed birds. This distribution is displayed in Figure 3.

Next, suppose De Roos inspects his population on three days providing the same results as on the 9th, 10th and 11th day of Table 1 such that $m = 14$. Using $k = 3$ and $q = 1 - (0.43)^3 = 0.92$ we recommend the distributional inference

$$\tilde{Q}(m) = 0.96\text{NegBin}(14, 0.92) + 0.04\text{NegBin}(15, 0.92)$$

for the frequency of unseen ringed birds (see Section 3). By shifting this over $m = 14$ units to the right, we obtain a distributional inference about r, see again Figure 3. Note the 'convergence' of these opinions to the true value of r, which we believe to be 14 or 15, or perhaps 16.

Acknowledgements

The comments by the referees led to improvements, especially in Sections 3, 5 and 6. Ornithologists R. H. Drent and J. B. Hulscher provided us with references and comments.

References

Conn, P. B., W. L. Kendall and M. D. Samuel (2004), A general model for the analysis of mark–resight, mark–recapture, and band-recovery data under tag loss. *Biometrics* **60**, 900–909.

de Roos, G. Th. (1987), Wadvogelexpeditie naar Australië, *Waddenbulletin* **22**, 188–192.

Hempel, C. G. (1965), *Aspects of scientific explanation, and other essays in the philosophy of science*, Free Press, New York.

Jaynes, E. T. (2003), *Probability theory: the logic of science*, Cambridge University Press.

Kardaun, O. J. W. F. and W. Schaafsma (2003), *Distributional inference, towards a Bayes–Fisher–Neyman compromise* (available on request).

Kroese, A. H., E. A. van der Meulen, K. Poortema and W. Schaafsma (1995), Distributional inference. *Statistica Neerlandica* **49**, 63–82.

Otis, D. L., K. P. Burnham, G. C. White and D. R. Anderson (1978), *Statistical inference from capture data on closed animal populations*, volume 62, Wildlife Monographs.

Parzen, E. (1960), *Modern probability theory and its applications*, John Wiley & Sons.

Polya, G. (1975), Probabilities in proofreading, *American Mathematical Monthly* **83**, 45.

Salomé, D. (1998), *Statistical inference via fiducial methods*, PhD thesis, University of Groningen.

Stam, A. J. (1987), Statistical problems in ancient numismatics, *Statistica Neerlandica* **41**, 151–173.

White, G. C., D. R. Anderson, K. P. Burnham and D. L. Otis (1982), Capture–recapture and removal methods for sampling closed populations, Technical Report LA-8787-NERP, Los Alamos National Laboratory.

Yang, M. C. K., D. D. Wackerly and A. Rosalsky (1982), Optimal stopping rules in proofreading, *Journal of Applied Probability* **19**, 723–729.

Received: January 2003. Revised: February 2005.